Protein structure mining using a structural alphabet

M. Tyagi[1,§], A.G. de Brevern[2], N. Srinivasan[1,3] and B. Offmann[1,*]

[1]Laboratoire de Biochimie et Génétique Moléculaire, Bioinformatics Team, Université de La Réunion, BP 7151, 15 avenue René Cassin, 97715 Saint Denis Messag Cedex 09, La Réunion, France.

[2] INSERM UMR-S 726, Equipe de Bioinformatique et Génomique Moléculaire (EBGM), Université Paris 7 – Denis Diderot, case 7113, 2, place Jussieu, 75251 Paris Cedex 05, France.

[3]Molecular Biophysics Unit, Indian Institute of Science, Bangalore 560 012, India.

[§] Present Address: Computational Biology Branch, National Center for Biotechnology Information (NCBI), National Library of Medicine (NLM), 8600 Rockville Pike, Bethesda, MD 20894, USA.

[*] Corresponding Author

List of abbreviations :

Protein Blocks : PBs
Secondary structure elements : SSEs
Structural alphabet : SA
Local alignment : LA
Global alignment : GA
Protein Block Expert : PBE
Root Mean Square Deviation : *rmsd*
Extreme value distribution : EVD

Abstract

We present a comprehensive evaluation of a new structure mining method, called PB-ALIGN. It is based on the encoding of protein structure as 1D sequence of a combination of 16 short structural motifs or protein blocks (PBs). PBs are short motifs capable of representing most of the local structural features of a protein backbone. Using derived PB substitution matrix and simple dynamic programming algorithm, PB sequences are aligned the same way amino acid sequences to yield structure alignment. PBs are short motifs capable of representing most of the local structural features of a protein backbone. Alignment of these local features as sequence of symbols enables fast detection of structural similarities between two proteins. Ability of the method to characterize and align regions beyond regular secondary structures e.g. N and C caps of helix and loops connecting regular structures puts it a step ahead of existing methods which strongly rely on secondary structure elements (SSEs). PB-ALIGN achieved efficiency of 89% in extracting true fold from a large database of 7259 SCOP domains and was successful in 96% cases to identify true super family members. On comparison to 13 existing structure comparison/mining methods, PB-ALIGN emerged as the best on general ability test dataset and was at par with methods like YAKUSA and CE on non-trivial test dataset. Furthermore, the proposed method performed well when compared to flexible structure alignment method like FATCAT and outperforms in processing speed (less than 45 sec per database scan). This work also establishes a reliable cut-off value for the demarcation of similar folds. It finally shows that global alignment scores of unrelated structures using PBs follow an extreme value distribution. PB-ALIGN is freely available on web server called Protein Block Expert (PBE) at http://bioinformatics.univ-reunion.fr/PBE/.

Introduction

Protein Data Bank (PDB)[1] offers to date more than 43,000 protein structures in the public domain. This data encloses wealth of information, which plays a critical role in our understanding of protein function, its evolution and sequence to structure relationship that further unfolds improved solutions to structure prediction and validation. Mining of information from this huge amount of data plays a crucial role and most of the time information includes measure of structural similarities between two or more proteins. Steady increase in number of known proteins has made it impossible for manual inspection of each and every protein present in PDB with rare exception of SCOP[2] database based on manual classification of protein domains. To overcome this limitation many groups have developed various structure comparison methods[3,4].

Structure comparison between two proteins has been of major interest from the time when Perutz[5] used structure alignment to highlight structural similarities between myoglobin and hemoglobin despite sharing low sequence similarities. Common functionality between these two proteins is not accidental, evolutionary relationship and shared structural features are the reasons that describe above phenomena. Structure comparison methods are aimed to find these structural similarities to foresee protein's function specially at low sequence similarity, to study evolutionary relationship and basic understanding of protein folding problem.

Furthermore, comparison and alignment helps in organization and classification of known proteins[6,7], mining of similar known proteins for newly solved structure[8-11], identification of functionally important sequence pattern in homologous proteins[12,13] and also provides point of reference for sequence alignment methods[14-16]. Structure comparison and alignment is more challenging and complicated problem due to number of reasons. First problem is about what to compare? Once this decision is made it is very difficult to obtain optimal alignment or to identify among many alignments as there are number of ways to align two structures. Also, in presence of high structural similarity it is

challenging to make out if it arises from evolutionary constraint or is just an analogy due to physical constraint on fold space.

There are various methods for structure comparison based on what level structure is represented; some methods use all atom model but are limited to small substructures[17]; more common approach involves backbone comparison of two proteins based on backbone atoms e.g. Cα atom or internal distance matrices[9,18,19] or internal angles[10,20,21] estimated from backbone atoms. Alignment of structures based on initial alignment of secondary structure elements (SSE) and further refinement through iteration is also a commonly used approach[11,22,23]. Graph theory based SSE alignments provides another alternative solution for structure comparison[24,25]. Methods using backbone coordinates for structure comparison relies on root mean square deviation measure between two proteins and objective function has to minimize this value to identify structural similarities. Such methods[26-28] are useful for comparing two proteins or substructures but they are computationally expensive, making them too slow for mining of similar structures from large database. Recently developed methods like FlexProt[29] and FATCAT[30] use flexible structure alignment approach by introducing twists between aligned fragment pairs to improve overall superposition and try to overcome the limitations of rigid body structure alignment techniques.

Popular methods like DALI[9], SSAP[19] and CE[18] use reduced representation of protein backbone in terms of distance matrices. DALI uses hexapeptide distance matrices combined with dynamic programming and Monte Carlo optimization technique to obtain global alignment. The Combinatorial Extension (CE) method combines aligned short structural fragments into larger alignment paths and apply dynamic programming to generate global alignment. Both these methods are most commonly used for structure comparison and fast structure mining though sometimes absence of homologue in database can increase search time considerably.

Most of these methods perform structure alignment based on secondary structure elements (SSE) or use them to obtain initial starting point. Protein structures can also be

approximately described using structural alphabets (SA), which are recurring short structural motifs found across protein 3D space (for a review, see Offmann et al[31]). Many groups have identified these recurring short motifs capable of describing protein backbone[32-34] and are believed to be more informative in protein structure analysis[35].

Use of structural alphabets for structure comparison has been attempted only in the last decade. 3D-Blast is an example of one such recently developed approach, which uses a 23 states structural alphabet to describe the backbone[36]. This method, uses BLAST as a search method using a structural alphabet substitution matrix to find the longest common substructures with high-scoring segment pairs. Though this method uses an E-value as measure of statistical significance of an alignment and generates results with performance comparable to known methods, the authors still did not assess the confidence index of the methodology.

Using a set of 16 pentapeptide structural motifs known as protein blocks (PBs)[37,38] we have introduced a new methodology of analyzing protein structures[39]. Each of these 16 motifs are represented by character alphabet (*a, b, c,...p)* and are described by vector of 8 dihedral angles (ϕ, ψ) making it possible to represent 3D protein structure by a string of 1D sequence of PBs. Taking advantage of this reduced representation of protein structure as mere sequence of symbols, we recently derived a PB substitution matrix and investigated its potential utility in protein structure analysis[31,39] or for discovering functional local structural motifs[40].

A new structure comparison method (PB-ALIGN) useful for mining protein structural databases has been developed. This approach is based on PB sequence alignment using the newly derived PB substitution matrix which has been developed[39]. The basic premise of structure alignment is very simple and is based on encoding of protein backbone by a sequence of characters representing PBs. Further; these PB sequences are aligned just like amino acid sequences using dynamic programming combined with a substitution matrix. Capability of PBs to represent local structure variations and alignment of these PBs provides more intuitive knowledge of structurally similar regions in two proteins when compared to SSE representation. Structure

alignment based on PB sequence is not only able to align regular substructures but also N and C cap regions. It also highlights structural variations in loops that connect regular secondary structures. PB sequence alignment to obtain structure alignment is a very fast procedure of structure mining and allows large database mining in real time[41].

In the present study, we provide a comprehensive evaluation of the methodology compared to existing techniques. We also provide more thorough analysis of the efficiency rate of mining proteins from a large database, using PBE server. In addition we present optimal gap penalty for both local[42] and global[43] alignment techniques. Our results show PB-ALIGN provides equivalent or better efficiency rate in mining of structures in from large database when compared to methods like DALI[9], CE[18] and FATCAT[30] and is much faster in all of them. Our method achieved 89% success rate in extracting true fold from pairwise alignment of 7259 against 7259 SCOP domains. In response to difficult test cases PB-ALIGN provides comparable results to more robust and complex methods and also gave satisfactory results while handling multi domain proteins. We addressed the question of alignment score threshold for making decision that two aligned structures correspond to same fold. The statistical characteristics of the distribution of global alignment scores were finally examined.

Materials and methods

Data set used for evaluation of PB-ALIGN

In the present study we have used batteries of test dataset to assess the performance PB-ALIGN in different experimental conditions. Database of 7259 SCOP (v1.65) domains filtered at 95% identity implemented in PBE web server[8] are used for assessing the mining efficiency of the method. Distribution of seven SCOP classes is as following: 1337 (18.5%) alpha domains, 2077 (28.6%) beta domains, 1387 (19.0%) alpha beta domains, 1529 (21.0%) alpha plus beta domains, 700 (9.6%) small domains, 89 (1.2%) multi-domain, and 140 (1.9%) membrane domains. PB-ALIGN was compared with 13 existing structure mining/comparison methods based on three different datasets. The general ability of the methods to extract similar structure proteins was tested on 61

query proteins belonging to ten protein families, representing the four CATH main classes (mainly α, mainly b, mixed α–β and few secondary structures). Same dataset was used in two independent studies done by Novotny et al[3] and Carpentier et al[10]. Ability of the methods to handle multi domain proteins was evaluated based on two multi domain queries selected by same groups. 14 non-trivial query-target pairs were taken from study done by Carpentier et al[10] to test the robustness of the method in detecting difficult structural similarities. Furthermore, we performed comparison of PB-ALIGN with flexible structure alignment program FATCAT, based on pairwise alignment of 10 difficult pairs as used by Ye et al[30].

Encoding 3D structures into PB sequence

Local backbone features of a protein can be represented by 16 prototypes of five residue long motifs called PBs[38]. Each PB is characterized by vector of eight dihedral (ϕ,ψ) angles associated with five consecutive Cα atoms and the 16 PBs are denoted by a character set varying from *a* to *p*. Encoding of protein backbone into PB sequence is a two step process; (i) coordinates of backbone atoms are used to calculate sequence of (ϕ,ψ) angles, (ii) An overlapping window of eight (ϕ,ψ) angles (corresponding to five Ca residues) is moved along the backbone. PBs for each window is assigned on the basis of smallest dissimilarity measure called root mean square deviation on angular values or *rmsda* [44] calculated between observed (ϕ,ψ) values in the window and the standard dihedral angles for various PBs. By following the above simple procedure a 3D structure of a protein can be encoded into a 1D sequence of PBs representing local structural information as sequence of structural alphabets.

PB substitution matrix

A 16 x 16 PB substitution matrix has been recently derived by our group[45]. The substitution scores between PBs were evaluated by counting the number of substitutions occurring in conserved regions of structurally aligned homologous proteins. These proteins are selected from large database, PALI[46,47] containing structure-based pairwise and multiple alignments of homologous proteins of known three-dimensional structures. The database uses a rigid-body superposition program, STAMP[48] to generate structure

based alignments. In total 21,503 pairwise alignments from 1197 SCOP families were analyzed which accounted for more than 2,000,000 PB substitutions. The raw frequencies are normalized and expressed as the log-odds score. The obtained scores provide extent of preference of a PB in a protein for its retention or substitution and allow to evaluate equivalence between homologous structures. The matrix has been validated in our previous studies and has been shown to be useful in identification of structurally equivalent regions in two proteins. In addition, the matrix has potential applications in differentiating between conformational differences and rigid body shifts among homologous protein structures[45].

Gap penalty optimization

In our previous study we selected arbitrarily gap penalty of –0.5 on manual inspection of PB alignments. Here we follow extensive procedure to suggest optimal gap penalties. Penalty optimization procedure is based on two criteria; effect of gap penalty on overall mining efficiency of similar structure proteins and quality of alignments generated. Structure mining efficiency is measured by counting number of times a true hit at class, fold, super-family and family level is obtained when top 10, 5 and first ranking alignments are considered for a given query. Quality of alignment is measured in terms of *rmsd* value obtained from superimposition of protein pairs based on PB alignment. Superimposition is performed using ProFit[49] software where equivalent zones are specified by PB alignment.

We performed a comprehensive study to suggest optimal gap penalty for both local alignment (LA) and global alignment (GA) algorithms using 2000 randomly sampled domains. A database of 2000 x 2000 pair-wise PB alignments was generated to perform above two analysis. Attention was given to keep the relative proportion of seven major classes similar to as in original databank. Jackknife approach was used to measure mining efficiency and alignment quality measure was done by considering only pairs belonging to same family. For global alignment of PB sequences we used following set of gap penalties –0.5, -2.0, -2.5, -3.0, -5.0 and optimal gap penalty for LA algorithm was selected from following set of penalties -0.5, -2.0, -3.0, -5.0 and –7.0.

Results and discussion

Effect of gap penalty on mining similar protein structures

Using 2000 randomly sampled domains, we assessed efficiency of both local and global alignment techniques to extract structurally similar proteins at class, fold, super-family and family levels for a given gap penalty. For a given query, hits are calculated by considering top 10, 5 and first ranking alignments. In the following analysis we present results from top 10 ranking alignments.

Table 1 reports efficiency rate for mining proteins at class, fold, super-family and family level considering top 10 ranking alignments based on GA algorithm. Bold values indicate best efficiency rate achieved at each level. With varying gap penalty from –0.5 to –3.0 negligible effect on efficiency of extraction of proteins was detected. Not more than 0.6% of change in seen in efficiency rate in this range of gap penalties. Further increase reduces the efficiency of the method by almost 3% as illustrated from low success rate achieved at –5.0 penalty. Among penalties used in this analysis, gap penalty of –2.0 seems gave best results though performance was not very much higher from other penalties used.

Table 2 shows the success rate of mining similar proteins using LA at class, fold, super-family and family level when top 10 ranking alignments are taken into account. Due to the basic nature of the algorithm it was suspected higher gap penalty will yield better results and efficiency rate can be inferior to GA. Indeed the two assumptions are true from the above tables. Efficiency rate has increased by almost 10% at fold level by changing penalty from –0.5 to –2.0 but overall success rate is slightly lower when compared to global alignment results. One of the reasons for low efficiency compared to GA can be due to the dataset used in our analysis. Since we have taken well defined domains as query against well defined domain database hence global alignment has an advantage here due to basic nature of the algorithm. This advantage can be limiting factor for GA in case we use complete protein chains as query without any knowledge of domain boundary. Indeed this is further documented in the following sections where LA

out performs in real case scenario and is able to extract true domains with high scores from the database whereas GA fails due to more number of gaps introduced in the alignment. On varying penalty from –2.0 to –7.0, the variation on efficiency rate is very moderate though best results are obtained at –3.0 or –5.0 gap penalty.

Similar efficiency rates achieved by neighboring gap penalties, both in local and global alignment techniques indicate variation in gap penalty increases or decreases success rate only to certain extent. From the previous analyses, no clear favorable gap penalty can be considered as optimal. From manual inspection of alignments obtained from various penalties indicated even though mining rate is somewhat constant, gap penalty can have more impact on quality of alignment produced. Based on this assumption, we further studied the relationship between gap penalty and quality of alignment in the following section.

Effect of gap penalty on structural alignment quality

In this analysis we performed a very simple exercise whereby, for each gap penalty, we generated both local and global alignments between pairs of homologous structures belonging to the same family. To assess the quality of PB based alignments, we used *rmsd* values from the superimposition of aligned residues. Each PB alignment was converted into corresponding amino acid (AA) alignment and was presented to ProFit software, which further performed least square fit of backbones based on AA alignment. List of *rmsd* values for every pair was complied at different gap penalties. For comparison of overall effect of gap penalty on *rmsd* values, we plotted the average improvement in *rmsd* values on different gap penalties with respect to *rmsd* values obtained at penalty of –0.5. Basically we have tried to highlight the change (decrease) in *rmsd* values of various pairs at different gap penalties when compared to values obtained at gap penalty of –0.5. Figure 1a shows the increase in improvement of average *rmsd* values at different gap penalties (–2.0, -2.5, -3.0 and –5.0) for GA algorithm. From the above figure it is very clear that increase in negative gap penalty has shifted more number of pairs towards lower *rmsd* values. For example, on varying gap penalty from –0.5 to –

3.0 almost 18% of alignment pairs have lower *rmsd* values in interval of 0.5Å to 1Å and gap penalty of –3.0 and –5.0 has brought improvement of 1Å or more to almost 44% and 47% homologous pairs respectively (data not shown). Figure 1b shows similar improvement in local alignment quality by varying gap penalty from –0.5 to –7.0. Once again overall improvement in *rmsd* values is seen at various gap penalties. Most fruitful penalties were –5.0 and –7.0 where average improvement of more than 2 Å is observed.

Based on efficiency of mining similar proteins and improvement in quality of alignment at various penalties level it was found –3.0 and –5.0 were optimal gap penalties for global and local alignment algorithm respectively. For GA penalty of –2.0 yielded best results for extraction of similar proteins but gap penalty of –3.0 was optimal value in terms of overall alignment quality and mining rate. In case of LA algorithm even though –7.0 gap penalty was able to give better *rmsd* values the extraction rate was inferior to –5.0 penalty by almost 1% and hence –5.0 was chosen as optimal penalty having balanced results for both alignment quality and mining of proteins.

Mining of protein structures

Efficiency of PB-ALIGN method to extract structurally similar proteins at different SCOP classification level was tested in the following study. We have analyzed 7251 domains selected from SCOP data bank filtered at 95% identity and performed all against all pairwise global alignment of PB sequences with an optimized gap penalty of –3.0. A class confusion matrix has been generated to from 7259 x 7259 pair-wise alignments to assess discriminatory power of simple PB alignments to assign correct SCOP class. The method was evaluated by counting if the true class, fold, superfamily or family member is present within top 10 hits, ranked by normalized score. It is noteworthy that performance of PB-ALIGN was evaluated in a jack-knife / leave-out approach where each query domain was removed from database prior to testing. This corresponds to real-life situations when one will have to query structural databases for mining similar folds and to what is current practice in published studies for evaluation performance of structure mining methods[3,4,9,10,50]. However, so as to evaluate efficiency of PB-ALIGN to assign high classification levels and capture remote homology, we also evaluated the

situation where members from same family as query was removed. So the dataset size used in all cases dynamically changed depending upon the query protein.

Table 3 summarizes the results obtained for each level, mainly class, fold, superfamily but also family at three different ranks, 1st, 5th and 10th. True class of a query protein can be found with an efficiency rate ranging between 92.5% and 98.2% when first 10 ranked alignments were considered. Possibility of finding true fold among Top10 hits showed distinct performance with a hit rate of 65.1% when whole family is jack-knifed and 87.4% when only query is left-out and similar performance were observed at superfamily level. Taking into account only the first hit (Top1 column in Table 3) we found between 76.1 and 93.1% success in finding true class. Among these first ranked hits, about 81.3% was from same fold when only query was jack-knifed but of 53% when whole family was removed. Similar performance was obtained at superfamily level (47.4-79%). Finally, on average, one is able to find true family of protein in 95% of cases.

It is further evidenced from our results that, when higher level is properly identified, the chance to identify subsequent lower level is very good. For example, from Table 4, when whole family is removed, and considering Top1 hits, out of a total of 3846 queries that were properly assigned at the fold level, 3438 (i.e 89.%) were also correctly predicted in their corresponding superfamilies. PB-ALIGN hence features robust nature in locating super-family relationships (functional relationships). The biggest decrease in prediction efficiency is observed between class and fold levels. However, it is noteworthy that because we look only at Top10 or Top 1 for performance evaluation, in some instances, for e.g. the query d1g73a_ from scop family a.7.4.1, the good hit is found only at lower rank, here at 69th rank because top scores were populated with redundant hits from homologous members of other folds (e.g. 23 hits from a.1.1.2. family, 14 hits from a.1.1.3 family, etc.).

These overall results indicate that mining similar structures using simple PB alignment methodology is efficient for identifying class and super-family relationships despite the presence of confusion across fold alignments. Indeed, when whole family was left-out, though it performed reasonably well, structural relationship at fold level was more difficult to capture using PB-ALIGN when compared to other structural levels. This highlights the importance of quality of information and its coverage in the databases used

for mining structures. However, in real life situations, users of structure mining tools such as PB-ALIGN generally expect that their structures be compared to whole PDB or representatives from all families. Hence, overall result presented in Table 3 is useful to assess how the method is performing when a query has a counterpart from the same family in the PB-ALGN database. It also shows that the method is able to capture higher-levels relationship (superfamily or fold) when the query has no homologues in the database. These encouraging results demonstrate the feasibility of the method to be useful in projects like structural genomics.

What Table 3 does not tell us is how much confusion exists between SCOP classes due to reduced representation of 3D structure using PB alignment. In order to address this question, a class confusion matrix was generated, using a 7259 x 7259 pairwise alignment, as shown in Table 4. It is important to analyze this confusion because by using 1D PB representation some topological information maybe lost and it becomes crucial for proteins sharing similar succession of secondary structure elements (SSE). Again both situations where query-only or whole family related to query is removed from the database were analyzed.

As shown in Table 4, beta class was most efficient (89.3 - 96.5%) to identify itself, closely followed by alpha and alpha beta (AB) class which have efficiency rates of 83.7 - 95% and 83.8 - 95.7% respectively. Alpha plus beta (AplusB) class was found to be confused with other classes with a 88.6% success rate when only query was left-out and a 56.4% rate when family was jack-knifed. Almost half of the false hits from Alpha plus beta are confused with Alpha-Beta class. Overlap between AB and AplusB is understandable taking into the fact that both have successions of helical and sheet regions. Performance for identifying small proteins was equivalent to AplusB class with rates of 65.3 – 89.8%. Other two classes has very contrasting results based on the jack-knife procedure ; when whole family is removed, probability to get true class as Top1 hit drops from 72.8 – 78.6% to 20.2 – 37.8% for membrane and multidomain proteins. Multidomain proteins are mostly confused with Alpha-Beta class while membrane proteins, as expected, are mostly confused with Alpha class.

Computation of class prediction matrix i.e. confusion matrix using large number of domains highlights the efficiency of PB alignment. Good efficiency rate is an indication that reduced complexity of 3D space and absence of topological information in PB representation has not affected the discriminatory power of PB alignment. This efficiency level can be attributed to combination of PBs connecting similar SSEs in different topologies (see below).

Efficiency rate within SCOP classes

Each class was studied separately to quantify how success rate at fold, super family and family was distributed within each class. Table 5 gives success rate for seven major SCOP classes namely, alpha, beta, AB, AplusB, multi-domain, membrane and small proteins when whole family related to the query was removed or when the query only was jack-knifed.

Success of finding true fold among Top10 hits was best for beta and AB class with an efficiency of 71.1 - 93% and 64.9 - 92.2% respectively, followed by membrane (87.8 – 90.7%), small (66.7 - 89.4%), AplusB (66.5 - 87.5%) and alpha (62.3 - 86.7%) class proteins. Similar trend was followed at super family and at family level. Looking at hits that ranked first, beta and AB classes achieved 53.8 – 88.7% and 48.9 – 87.8% of success respectively in finding true fold compared to alpha class where only 56.6 – 77.1% efficiency was reached. Presence of long helical regions in alpha proteins can be one of the reasons for more confusion among various folds in alpha class (as illustrated in the following discussion). Interestingly, distribution of success rates at three SCOP levels very well indicate true hits at fold and super family levels are not only populated by family members. These results are on same line as observed in Table 3. Analysis of above results indicate PB alignment is able to locate structure similarities even at very low sequence similarities (superfamily relationships) and thus our method can be used to detect remote homologues for a given protein.

Closer look at the cases of failure where query protein was not able to find its true fold, super-family or family within top 10 ranks gave insight to current limitations of

structure mining. Presence of single member folds, superfamilies or families in database was most common contributor to absence of true hit. In some instances diversity within family both at in terms of length and structural features makes it difficult for the global alignment algorithm to extract true member. For example, ABC transporter protein from *Sulfolobus solfataricus* (SCOP domain d1oxsc1) from MOP-like super-family (SCOP code b.40.6.3) is almost 40 residue longer than rest of the members. In this case GA algorithm have to introduce large number of gaps to accommodate shorter proteins from the same family resulting in low alignment scores hence low rank. Application of LA provides an alternative solution in this case and enabled to extract at least one member protein in top hits. Human hyperplastic discs protein (SCOP domain d1i2ta_) from PABC (PABP) domain family (SCOP code a.144.1.1) is another such example where LA was able to extract true hit among top ranks and GA was unsuccessful. Note should be taken that not in all such cases success is achieved by LA approach. For example Haloarcula marismortui protein (SCOP domain d1jj2s) from Ribosomal proteins L24p and L21e family (SCOP code b.34.5.1) is one such example where both diversity in structural features and protein length plays a role in unsuccessful results. Such examples are challenging and provide an opportunity to refine and improve our approach.

Furthermore, PB alignment technique had some problems with the pair of proteins sharing long stretch of regular secondary structures e.g. long helices in alpha proteins. In PB sequence these regions are represented as long stretch of PB *m*'s, alignment of such regions artificially contributes to the global score and put them in high rank. This example is very well illustrated from pairwise alignment of ribosomal protein L12 from *Thermotoga maritima* (SCOP domain d1dd3a1) and ROP protein from *E. Coli* (SCOP domain d1b6q__), Figure 2. The figure shows good alignment of helical regions (sequence of PB *m*'s), major reason for having high alignment score. Closer look at alignment indicates presence of extra loop in domain d1dd3a1, which is absent in other protein. Detection of this extra loop in one protein can hint in the difference in relative orientations of helices in two proteins even though the alignment score is high and is evident from Figure 2. This example highlights sometimes alignment score can be misleading due to the high content of regular structures in two proteins but closer look

into PB alignment can give clues to the structural differences. In such cases disadvantage of 1D representation can be overcome by having a manual inspection to detect local variations present in between regular structures.

Comparison with existing methods

Performance of PB-ALIGN has been tested against 13 structure comparison methods, Table 6. We applied batteries of tests to PB-ALIGN to obtain a comprehensive comparison with existing methods which included general efficiency of the method to extract related proteins, ability to identify difficult structure similarities in database search, performance to handle multi-domain proteins and comparison with flexible structure alignment method like FATCAT. Evaluation results of existing methods are taken from recent studies done by Novotny et al.[3] and Carpentier et al.[10] where 12 structure comparison methods were evaluated. Comparison with flexible structure alignment method is based on results produced by Ye et al.[30] where 10 difficult pair alignments are compared between VAST, DALI, CE and FATCAT. Dataset of 61 queries to compare general ability to extract related proteins and 2 multi-domain protein queries are taken from study done by Novotny et al. We followed same evaluation procedure as done by Novotny et al. and Carpentier et al. with one basic difference, both the studies followed CATH[6] classification for test dataset and in our study we have used SCOP[2] classification for evaluation procedure. This has been done because PB-ALIGN uses database of SCOP domains filtered at 95% identity and many times there are differences in both classification schemes. A hit is counted as true hit if it belongs to same SCOP super-family or family level. It should be noted that some of the methods involved in this comparative study use significance threshold for the scores obtained to discriminate same fold from different folds but not all. However, our method is not initially based on the definition of a threshold measure and considers top hits in every mining exercise to evaluate the relative performance of methods. Nonetheless, measures have been developed to assess significance of the alignments (see paragraph cut-off below for details). Thus, we applied here similar protocols to Novotny et al that have used few hits to compare those methods which did not return the significance of hits[3].

From our initial tests we found out while using complete protein chain as query that LA algorithm gives far better results when compared to GA technique. This is obvious from the fact that many times protein chains are longer than actual domain boundaries and using GA in such cases increases number of gaps in alignment procedure. To avoid above pitfalls we have used LA for our analysis.

Using local alignment we queried each protein chain against the databank and complied results simply by counting if true hit is found within top 10 alignments, number of members found in top 10 and rank of 1st false positive. Alignments are ranked based on score generated by LA algorithm plus normalized and Z scores are also reported for each hit. Out of 61 queries we found 2 cases (1rlr and 1vmo) had no superfamily and family members in our databank except themselves. In total we tested 59 test cases for general efficiency of the method and compared our results with the results reported by Carpentier et al., Table 7. Overall PB-ALIGN performed with a success rate of 96.6% when top 10 ranking alignments are considered.

Our method performs correctly in all except two cases. Toxin protein 1ciy (PDB id) from *Bacillus thuringiensis* has three domains namely delta-Endotoxin C terminal, middle and N terminal domain. PB-ALIGN is able to extract C terminal and N terminal domains very easily in top hits but misses out target middle domain (SCOP code b.77.2.1) of 1ciy from initial hits and is found at only 48th rank. Due to low rank in hits we have counted this hit as negative in our final results. Second protein 1grj (PDB id) from *E. Coli* has two domains and target domain (GreA transcript cleavage protein, C terminal domain, SCOP code d.26.1.1) is present as lone member of family in our database. PB-ALIGN is not able to identify target family (single member family) or super-family members among top hits. Target super family members are obtained at 12th and 16th rank if mining is performed with gap penalty of –3.0. Our assessment also found out that for only 5 queries out of 59 tested, the rank of first false positive was above the rank of true hit and there was not a single case where first true positive was ranked after a false hit. Moreover time taken to query each protein was less than a minute, making it

one of the fastest and efficient structure comparison method among the 13 methods (see Table 7) evaluated in present study.

Performance of PB-ALIGN on nontrivial dataset

We further tested our method's performance on a non trivial dataset used by Carpentier et al.[10] which is constituted of 14 difficult cases. We queried each protein using both LA and GA algorithm against our database and considered first 100 alignments as well as a cut-off score value of -0.25 for global alignment (see below). Results of this exercise are comparable with previously data published by Carpentier et al[10]. PB-ALIGN performance was found to be at par with YAKUSA and CE and gave about 50% success rate with combined effort of LA and GA algorithm. Testing of both LA and GA approach on nontrivial dataset gave very interesting insight of our method. LA algorithm was able to get only 4 targets out of 14 tested while GA found total 6 targets with 3 extra hits from LA results and missing out one case found by LA. Table 8 gives the summary of results obtained using both the approaches. Application of GA not only identified more number of difficult targets but also improved the rank of target in two cases (Table 8). When cut-off value (see next section) is applied as a rule for decision, GA is able to capture 7 targets above the threshold value. Like YAKUSA, PB-ALIGN uses local structural features to describe and encode protein backbone and identifying distantly related proteins can be difficult due to the fact that such pairs share structural similarities at global level rather than local level. Ability of PB-ALIGN to use GA algorithm to capture such remote similarities at global level using local descriptors (PBs) sets it apart from methods like YAKUSA.

Test case 1crb and 1bge from mainly-α class demonstrate this fact very clearly by analyzing global alignment of query (1crb) and target (2gmf) protein PB sequences as shown in Figure 3a. Using GA we were able to find target for query protein 1crb among top hits with low alignment score. Looking at PB alignment it becomes very obvious why local alignment method was not able to capture these global similarities. Both the proteins share four helical regions (populated by PBs *m*) separated by loops and small strands (populated by PBs *c* & *d*) and two middle helical regions are of different length.

Such global similarity can only be captured by GA algorithm having lower gap penalty compared to high penalty imposed by LA approach. Use of simple GA algorithm combined with PB substitution table highlights subtle similarities identified by PB-ALIGN method based on local backbone descriptors (PBs). Figure 3b also shows a query protein (1bge) for which target protein (2gmf) was not found even after using GA. Close inspection of PB alignment reveals four helical regions are well aligned despite the differences in helix length. Presence of many gaps to obtain this alignment results in very low alignment score which pushed down the pair below the top hits. Another example where target (2fox) was found by using GA is query protein 3chy from mainly-αβ class, Figure 3c shows superimposition of 3chy and 2fox from ProFit based on global PB alignment. Once again identification of such similarities is not possible by using LA techniques due to the presence of variable regions and has to be accommodated by gaps in an alignment. The above results show that the possibility to align PB sequences using local or global alignment techniques offers flexibility to recognize both strong local similarity and distant (variable) global similarities shared by proteins.

Protein 2afn was the only query where LA outperformed GA. Prime reason of GA failure can be understood from the fact that query protein chain 2afnA contain multiple domains and global alignment against our database will need to introduce large number of gaps resulting in low alignment score. Whereas in case of LA algorithm only probe and target domains are aligned with high alignment score. These results indicate usage of GA algorithm on sequence of PBs (encoding local structure variations) is better suited to situations where structural similarities are shared at global level and are difficult to obtain with local alignment techniques. In situations, where protein chains are suspected to contain multiple domains or one protein structure is completely or partially contained in other protein, LA approach proves more advantageous.

Global alignment of few difficult pairs showed at global level that the method was able to align equivalent regions in two proteins but due to the low scores such pairs were missed altogether by database search approach. We extended above analysis by studying pair-wise alignment of tough cases and compared alignment results with flexible

alignment method called FATCAT. We selected 10 difficult pairs used by Ye et al and compared PB-ALIGN with VAST, DALI, CE and FATCAT based on number of residues aligned and superposition *rmsd* obtained. Table 9 shows number of residues aligned along with *rmsd* values (within brackets) based on GA of PB sequences. Other methods are compared based on the results obtained by Ye et al. In our case, superposition of two proteins was done using ProFit based on alignment provided by PB-ALIGN and further iterations where performed by ProFit to obtain final results. In other words, results presented in Table 9 are combined effort of PB-ALIGN and ProFit. To measure the contribution of ProFit we also performed another exercise where superposition was performed based on sequence alignment generated by ProFit to define initial equivalent zones and carried out iteration from there on to get final values.

Results we obtained are very interesting since, ProFit alone by itself gave comparable results to FATCAT and PB-ALIGN in 7 out of 10 cases. Remaining 3 pairs (1cewI, 1molA; 1cid_,2rhe_; 1crl_,1ede_) gave much improved results when combined with PB-ALIGN. This outcome has two implications; first, ProFit by itself is good enough superimposition method to superimpose protein pairs sharing low sequence similarity and can achieve comparable results to more complex and robust methods; second, in the cases where ProFit fails to find optimal results by simple amino acid sequence alignment, PB alignment provides good starting points to ProFit, unidentifiable by sequence alignment alone. In all ten cases PB-ALIGN coupled with ProFit gave desirable results compared to other complex methods. When compared to a flexible alignment and superimposition method FATCAT, PB-ALIGN gave low *rmsd* in most of the cases with slightly less number of residues superposed. It is noteworthy that methods like FACAT has real advantage in this study where it introduces twists in structures to superimpose more residues with low *rmsd* and despite this advantage our simple methodology gave decent results. The only test case (pair 1crl_, 1ede_) where PB-ALIGN produces significantly lower results compared to FATCAT can be understood from the fact that 5 twists were introduced in protein structure to superimpose 269 residues with a *rmsd* of 3.55Å. In its present form, though PB-ALIGN will align PB sequences in a flexible manner, it is still not capable to produce such results as it relies on

rigid body superposition method. Objective of this analysis was not to compete with methods like FATCAT (which we believe in principle will give better results specially in cases where twists are needed to superpose structures) but it is to highlight, (i) despite using very simple approach, PB alignment technique gives comparable results in most of the situations with minimum computation time making it practical for large scale analysis in real life situations and (ii) premises for flexible structural superimposition as performed by FATCAT are featured in the method of PB alignments owing to the nature of the algorithm used.

Handling of multi domain proteins

PB-ALIGN was also tested on two multi domain proteins, 2src_ and 2hckA (human Src and Hck kinase proteins respectively) to assess efficiency of method to handle proteins chains containing multiple domains during database search. The database used in our case is a collection of domains on SCOP classification, hence it was calculated whether the method is able to extract target domains among top hits. As seen above, LA technique has advantages over GA algorithm in such cases. Hence, in present analysis we used LA algorithm to extract different domains from database. Based on SCOP classification, query proteins are composed of three different domains namely SH3 domain, SH2 domain and protein kinases catalytic subunit. Our evaluation on multi domain proteins is slightly different from the earlier studies[3,10] where success was measured if hits contained all the four domains (based on CATH classification) followed by proteins having two or one domain. In our study we assessed if all the domains (SH3 domain, SH2 domain and protein kinases catalytic subunit based on SCOP definition) are present among top 100 hits. Previous studies reported YAHUSA, DALI, VAST, MATRAS and CE gave best results while handling multi domain cases. SSM found proteins having all 4 domains and TOP and DEJAVU found structures sharing more than one domain while having a blind eye to single domain structures. LOCK managed to find representative of each domain but failed to find proteins having all domains in single chain. TOPSCAN, PRIDE and TOPS were among least efficient methods. PB-ALIGN was able to find all three target domains among top hits. SH2 and kinases catalytic subunit domains where most easily found and were populated among top hits. SH3

domain was always found in at lower ranks (61st and 37th rank) and this can be attributed to smaller length and high population of other two domains in our database.

Cut-off threshold for PB_ALIGN scores to recognize common folds

On the basis of various assessment exercises described above we worked out a recommended threshold for the PB-ALIGN scores that will allow one to designate a hit in a structural database as same fold as that of the query. Figure 4 provides a distribution of scores of PB-ALIGN for the cases of common fold and different fold (according to SCOP). This Figure shows that a region of scores is taken-up by proteins with the same fold as well as different fold. As fold space is a continuum it will be difficult to have a precise score that will completely segregate same folds from different folds. We hence analyzed the variation of sensitivity and specificity for different normalized score thresholds (Figure 5). Because we want to typically minimize false positives while having an acceptable level of sensitivity (true positives), we propose to select appropriate cut off at a stringent specificity value of 0.95. Hence on the basis of the aforementioned specificity value, the normalized score cut-off value was -0.250 and the sensitivity, for proteins from the same fold was of 0.75. Similarly, for proteins from the same superfamily, the score cut-off value was -0.252 and a sensitivity of 0.87. Hence, we suggest a threshold value of -0.250 to discriminate between proteins from the same fold or same superfamily. This cut-off works correctly for demarcation of same folds or superfamily from different folds or superfamily although not for all the cases. Going by this argument cut-off, a total of 75% and 87% of the proteins with the same fold and superfamily as the query respectively are correctly picked-up by PB-ALIGN. Importantly, rate of false hits is only of 5% in both situations.

Understanding whether the observed PB sequence similarity is just a chance event is the central problem for the evaluation of the statistical significance of alignment scores. The basic question to be answered is: what is the probability that a similarity score as great as that actually observed between real sequences could have arisen by chance, when sampling from suitably-defined populations of unrelated sequences? In order to address

this question,the distribution of global alignment scores from real but unrelated sequences for different length subsets (40aa through 400aa by 30aa increment, see also Table 10) need to be analyzed. As shown for three examples in Figure 6, alignment scores were distributed according to an extreme value distribution (EVD). We hence derived all three corresponding EVD parameters (Table 10) which could be used to measure confidence of alignment scores to classify two proteins as having the same fold or belonging to same superfamily. The scale parameter (σ) linearly decreased with length alignment which further comforts the inferred EVD distribution model (Figure 7).

Conclusion

In the era of structural genomics, protein structure comparison and mining plays an important role in computational biology. Identification of new phylogenic relationships, functional annotation and study of sequence to structure relationships are some of its most common targets. In this study we have presented a new structure mining method called PB-ALIGN, based on the encoding of protein backbone as sequence of short local motifs (PBs) and their alignment using a newly derived PB substitution matrix and simple dynamic programming. The method is simple and is scalable for large scale analysis and provides an ideal choice of structural genomics. Use of local structural features (PBs) to describe protein backbone and alignment of such features provides an alternative to previously know methods like DALI and CE. Existing methods rely of SSEs information for structure alignment and misses out on large amount of structural information beside regular structures in proteins. PB representation of protein backbone highlights subtle variations and structural conservations present beyond local regular structures, e.g. N and C caps of helices and strands. Capability of PB-ALIGN to align these regions is a step ahead of existing methods.

The method is highly efficient in mining structurally similar proteins from large database at both fold and super-family level. Among peer comparison PB-ALIGN stood out as best in both efficiency to find target and speed to mine structures. Ability to obtain good efficiency at high speed highlights the simplicity and effectiveness of the method. Compared to methods like YAKUSA that also relies on representation of local structural

features our method performed superior on general test data and at par on difficult (nontrivial) dataset. The main difference between these two methods is the final objective. YAKUSA aims to locate strong gap-free local structural similarities or "blocks"[10] between two proteins and is not concerned with global similarities spread over protein length. Whereas, PB-ALIGN despite using local structural features aims to address both local and global similarity between proteins. Availability of PB substitution matrix and use of local or global sequence alignment techniques help to answer both local and global structure similarities. Both alignment techniques are shown to be useful in different conditions e.g. local alignment is beneficial if one wants to identify strong local similarities or if protein chain is multi-domain or if one protein is completely or partially included in other structure. Global alignment is useful if two proteins share structural similarities spread across protein length. Our experience suggests that user should use both local and global alignment feature and manual inspection of PB alignments will clearly highlight the best approach. Another advantage we found in PB-ALIGN is intuitive nature of PB alignment representing structure alignment and many times simple inspection of alignment gives hint about structural differences between proteins prior to 3D visualization.

Importantly, this study derived a score cut-off, for the inference of structural similarity between structural domains whose relationship is unknown using their PB representations. It further specified the extreme value distribution of global alignment scores of real but unrelated PB sequences. This information is currently being used to implement a statistical significance threshold in PB-ALIGN.

Comprehensive assessment of our methodology also highlighted some shortcomings and need of further fine-tuning of PB-ALIGN. At present we have used simple dynamic programming and linear gap penalty. We believe use of more robust flavor of dynamic programming and change in gap penalty will further improve the alignment quality. Artificial increase in alignment score due to long stretch of regular structures specially in alpha class proteins is also being looked into and change in scoring function is anticipated. Furthermore we would like to introduce combined scoring

function taking into account number of aligned residues, *rmsd* value and alignment score. We believe that the use of PB alignment methodology to perform multiple alignment of family members would enable use to define 'core' structures having boundaries beyond SSEs and would help in finding distant homologues. PB-ALIGN is also expected to be useful in homology modeling and loop modeling.

Tables

Table 1. Optimization of global gap penalty.

Effect of gap penalty on mining rate at class, fold, super-family and family level for global alignment. The results are from top 10 ranking alignments. Analysis was performed on 2000 randomly selected SCOP domains.

Level/Gap penalty	*-0.5*	*-1.0*	*-2.0*	*-2.5*	*-3.0*	*-5.0*
Class	**98.1**	97.9	97.9	97.95	**98.05**	97.9
Fold	66.35	66.55	**66.9**	66.5	66.25	63.85
Super family	61.5	61.65	**61.65**	61.35	61.05	58.7
Family	55.5	55.95	**56.6**	56.6	56.2	54.45

Table 2. Optimization of local gap penalty.

Effect of penalty on mining rate at class, fold, super-family and family level. Results are from top 10 ranking alignments. Analysis was performed on 2000 randomly selected SCOP domains.

Level/Gap penalty	*-0.5*	*-2.0*	*-3.0*	*-5.0*	*-7.0*
Class	89.9	93.3	93.9	94.2	94.25
Fold	50.55	60.55	62.9	62.75	61.35
Super family	49.15	58.15	60	60	59.05
Family	44.45	52.95	54.1	53.95	52.8

Table 3. Efficiency rate of mining proteins at various SCOP classification levels.

Results are reported for top 10 , 5 and 1st ranking alignments. Success rate at fold, super-family and family level was calculated only for those queries that were correctly predicted in class, fold and super-family level respectively**.** Two situations were distinguished ; one where only the query is removed from the database and another where the whole family was removed from the database. Values are given as percentage.

	Only query domain is removed from database			Whole family related to query is removed from database		
SCOP level	Top10	Top5	Top1	Top10	Top5	Top1
Class	99.1 (7194)	96.8 (7028)	93.1 (6758)	92.5 (6716)	88 (6394)	76.1 (5529)
Fold	87.4 (6343)	85.6 (6217)	81.3 (5906)	65.1 (4727)	60.8 (4412)	53.0 (3846)
Super Family	84.3 (6122)	82.8 (6011)	79.0 (5739)	62.6 (4548)	57.5 (4178)	47.4 (3438)
Family	**80.0 (5809)**	**78.7 (5714)**	**75.1 (5453)**	n/a	n/a	n/a

n/a : not applicable

Table 4. Class confusion matrix.

Matrix gives the efficiency of the method to find true class at first rank and the confusion rate between SCOP classes. Results were generated from 7259 X 7259 pairwise PB alignments. True classes are featured horizontally and predicted classes vertically. Two situations were distinguished within each class ; one where only the query is removed from the database (top line) and another where the whole family was removed from the database (bottom shadowed line).

True class vs. hit class	ALPHA	BETA	AB	APLUSB	MULTIDOM*	MEMBRANE	SMALL	Total
ALPHA	1271 (95.0%)	1	12	12	0	35	6	1337
	1120 (83.7%)	5	47	53	4	88	20	
BETA	2	2005 (96.5%)	10	36	0	3	21	2077
	3	1855 (89.3%)	20	120	2	18	59	
AB	7	8	1328 (95.7%)	39	3	0	2	1387
	31	20	1163 (83.8%)	145	20	3	5	
APLUSB	34	40	77	1356 (88.6%)	4	3	15	1529
	99	189	301	863 (56.4%)	22	4	50	
MULTIDOM	3	2	11	2	70 (78.6%)	0	1	89
	6	5	50	9	18 (20.2%)	0	1	
MEMBRANE	29	6	0	2	0	102 (72.8%)	1	140
	64	17	0	4	0	53 (37.8%)	2	
SMALL	23	24	3	20	0	1	629 (89.8%)	700
	53	115	5	69	0	1	457 (65.3%)	
								7259

*MULTIDOM corresponds to multi-domain protein class.

Table 5. Efficiency rate of mining similar structure proteins with in each SCOP class.

Efficiency is calculated at three different ranks top 10 hits, top 5 hits and 1st hit. Within bracket figures show the total number of queries taken into account or number of true hits. E.g. small (700) means in total we did this exercise for 700 proteins domains. Efficiency presented is in percentage i.e. true hits/(true hit + false hit).

	Only query domain was removed from database			**Whole family related to query was removed from database**		
Alpha (1337)	**Top10**	**Top5**	**Top1**	**Top10**	**Top5**	**Top1**
Fold	86.7 (1160)	84.3 (1128)	77.1 (1031)	69.8 (933)	64.4 (862)	56.6 (757)
Super family	83.2 (1113)	81.4 (1089)	74.4 (995)	62.3 (833)	56.0 (749)	45.2 (604)
Family	77.4 (1035)	75.0 (1002)	67.0 (922)	n/a	n/a	n/a
Beta (2077)						
Fold	93.0 (1931)	91.8 (1907)	88.7 (1842)	71.1 (1478)	66.3 (1378)	53.8 (1118)
Super family	90.0 (1869)	88.8 (1844)	85.8 (1782)	70.5 (1464)	65.1 (1353)	53.3 (1108)
Family	87.2 (1812)	86.0 (1786)	83.1 (1726)	n/a	n/a	n/a
AB (1387)						
Fold	92.2 (1279)	91.2 (1265)	87.8 (1218)	64.9 (901)	58.5 (812)	48.9 (679)
Super family	90.0 (1248)	89.0 (1234)	86.1 (1195)	63.4 (880)	57.0 (791)	48.8 (678)
Family	83.3 (1156)	82.6 (1146)	80.0 (1110)	n/a	n/a	n/a
AplusB (1529)						
Fold	87.5 (1338)	85.7 (1310)	81.9 (1253)	69.8 (1068)	65.0 (995)	59.7 (913)
Super family	84.3 (1290)	82.6 (1264)	79.2 (1211)	66.5 (1017)	59.9 (916)	51.5 (788)
Family	79.8 (1220)	78.3 (1198)	74.9 (1145)	n/a	n/a	n/a
Small (700)						
Fold	89.4(626)	85.8 (601)	73.8 (517)	66.7 (467)	61.7 (432)	47.1 (330)
Super family	84.7 (593)	80.8 (566)	70.6 (494)	61.5 (431)	58.2 (408)	49.0 (343)
Family	79.4 (556)	76.8 (538)	66.4 (465)	n/a	n/a	n/a
MultiDo (89)						
Fold	86.5 (77)	86.5 (77)	85.3 (76)	66.3 (59)	66.3 (59)	64.0 (57)
Super family	86.5 (77)	86.5 (77)	86.5 (76)	66.3 (59)	66.3 (59)	64.0 (57)
Family	76.4 (68)	76.4 (68)	76.4 (68)	n/a	n/a	n/a
Membrane (140)						
Fold	90.7 (127)	87.8 (123)	80 (112)	88.5 (124)	88.5 (124)	85.7 (120)
Super family	75.7 (106)	73.5 (103)	69.3 (97)	87.8 (123)	85.0 (119)	67.8 (95)
Family	73.5 (103)	71.4 (100)	67.8 (95)	n/a	n/a	n/a

n/a : not applicable

Table 6. Structure mining/comparison methods tested in the present study.

Program	URL	Methodology used
CE	http://cl.sdsc.edu/ce.html	Inter residue distances
DALI	http://www.ebi.ac.uk/dali	Cα distance matrices
DEJAVU	http://xray.bmc.uu.se/usf/dejavu.html	SSEs comparison
FATCAT	http://fatcat.ljcrf.edu/	*RMSD* and introduction of twists
LOCK	http://brutlag.stanford.edu/lock2	*RMSD* minimization
MATRAS	http://biunit.aist-nara.ac.jp/matras	Markov transition model
PB-ALIGN	http://bioinformatics.univ-reunion.fr/PBE/PBE-ALIGN.htm	PBs substitution matrix & alignment.
PRIDE	http://hydra.icgeb.trieste.it/pride	Cα distance distribution.
SSM	http://www.ebi.ac.uk/msd-srv/ssm	SSEs vector comparison
TOP*	http://bioinfol.mbfys.lu.se/top	SSEs alignments
TOPS*		SSEs symbolic representation and comparison
TOPSCAN	http://www.bioinf.org.uk/topscan	SSE representation in topology strings, aligned through a global dynamic alignment algorithm
VAST	http://www.ncbi.nlm.nih.gov/Structure/VAST/vastsearch.html	SSEs graph representation
YAKUSA	http://www.rpbs.jussieu.fr/yakusa	Internal coordinates matching

*Web link unreachable.

Table 7. Comparison of PB-ALIGN with existing structure mining/comparison methods.

Comparison of PB-ALIGN with 12 structure mining/comparison methods based on results from Carpentier et al.[10]. The numbers along with the header gives total number of queries belonging to each class. All the hits are counted based on first 10 ranking alignments compared to 100 hits taken by Carpentier et al, only for those methods which did not return the significance of hits.

Program	Mainly a (19)	Mainly b (19)	Mixed ab (15)	Few SSEs (8)	Total (%)
PB-ALIGN	18[+]	17[*]	14	8	96.6
YAKUSA	17	19	14	8	95
CE	17	19	13	8	93
DALI	14	19	14	8	90
MATRAS	11	19	14	8	85
VAST	12	17	15	7	84
TOP	14	18	12	7	84
DEJAVU	14	19	9	4	75
TOPSCAN	15	12	9	7	70
TOPS	2	15	14	7	62
PRIDE	14	14	7	3	62
LOCK	0	14	11	8	54
SSM	5	13	10	5	54

[+] One query has no target in our database. [*] For mainly b class, query protein 1vmo has no target in our database and query 1ciy misses target in top ten ranks.

Table 8. Performance of local and global alignment algorithm on nontrivial dataset.

In total both methods where able to find 7 target proteins. Success and failure of GA and LA is indicated by 0 (failure) and 1 (success). Here, target rank is indicated within parenthesis. Success of GA based on the application of a cut-off value of -0.25 is also indicated by 0 if score < -0.25 (failure) and 1 if score > -0.25 (success). Between parenthesis are indicated the normalized alignment score (see text for details).

Query protein	Target protein	Local alignment	Global alignment	Cut-off value on GA alignment
1aep	256b:A	0	0	0 (-0.29)
2mta:C	1ycc	0	0	0 (-0.74)
1rcb	2gmf:A	0	1 (48)$^{+}$	1 (+0.02)
1bge:B	2gmf:A	0	0	0 (-0.34)
2afn:A	1aoz:A	1 (10)*	0	0 (-0.48)
3hla:B	2rhe	0	0	1 (+0.23)
2aza:a	1paz	0	0	0 (-0.29)
1cew:I	1mol:A	0	1 (59)$^{+}$	1 (-0.06)
1dsb	2trx:A	0	0	0 (-0.68)
1fxi:A	1ubq	1 (42)	1 (28)$^{\#}$	1 (+0.26)
3chy	2fox	0	1 (33)$^{+}$	1 (+0.18)
1gpl	2trx:A	0	0	0 (-1.94)
1hip	2hip:A	1 (6)	1 (7)	1 (+0.56)
1isu:A	2hip:A	1 (59)	1 (15)$^{\#}$	1 (+0.14)

$^{+}$ Target exclusively found by GA. $^{\#}$ Improvement in rank using GA. * Target protein missed by GA.

Table 9. Comparison of PB-ALIGN and FATCAT.

Comparison of various methods based on number of residues aligned and *rmsd* (within brackets) obtained for 10 difficult examples based on global alignment. Results for other methods is taken from Ye et al.[30]

Protein1	Protein2	VAST	DALI	CE	FACAT	PBALIGN	ProFit
1fxiA	1ubq_	48(2.1)	60(2.6)[+]	64 (3.8)[+]	63(3.01)	59(2.6)	55(2.0)
1ten_	3hhrB	78(1.6)	86(1.9)	87 (1.9)	87(1.9)	82(4.1)	84(4.0)
3hlaB	2rhe_	-	63(2.5)	85(3.5)	79(2.81)[*]	67(2.4)	73(2.6)
2azaA	1paz_	74(2.2)	81(2.5)[+]	85(2.9)	87(3.01)	79(2.3)	79(1.9)
1cewI	1molA	71(1.9)	81(2.3)	69(1.9)	83(2.44)	74(2.5)	16(2.5)
1cid_	2rhe_	85(2.2)	95(3.3)	94(2.7)	100(3.11)	87(2.2)	25(2.9)
1crl_	1ede_	-	211(3.4)	187(3.2)	269(3.55)[*]	179(2.3)	75(2.9)
2sim_	1nsbA	284(3.8)	286 (3.8)	264(3.0)	286(3.07)	262(2.4)	264(2.4)
1bgeB	2gmfA	74(2.5)	98(3.5)	94(4.1)	100(3.19)	90(2.4)	88(2.4)
1tie_	4fgf_	82 (1.7)	108 (2.0)	116(2.9)	117(3.05)	105(2.2)	104(2.1)

[*] FATCAT introduced twists in structures to perform superposition. [+] No results obtained in previous study done by Ye et al.

Table 10. Estimates of extreme value distribution parameters for alignment scores between real but unrelated sequences of different sequence length subsets (see also figure 8). These parameters were derived using *gev* function in *evir* package implemented in R statistical software[51].

Length (number of residues)	Shape parameter (ξ)	Scale parameter (σ)	Location parameter (μ)
40	0.376	0.673	0.993
70	0.359	0.539	0.831
100	0.368	0.506	0.774
130	0.349	0.418	0.685
160	0.397	0.353	0.621
190	0.406	0.384	0.644
220	0.416	0.354	0.599
250	0.424	0.358	0.535
280	0.444	0.337	0.480
310	0.402	0.261	0.519
340	0.435	0.278	0.540
370	0.456	0.262	0.520
400	0.477	0.242	0.499

Figures

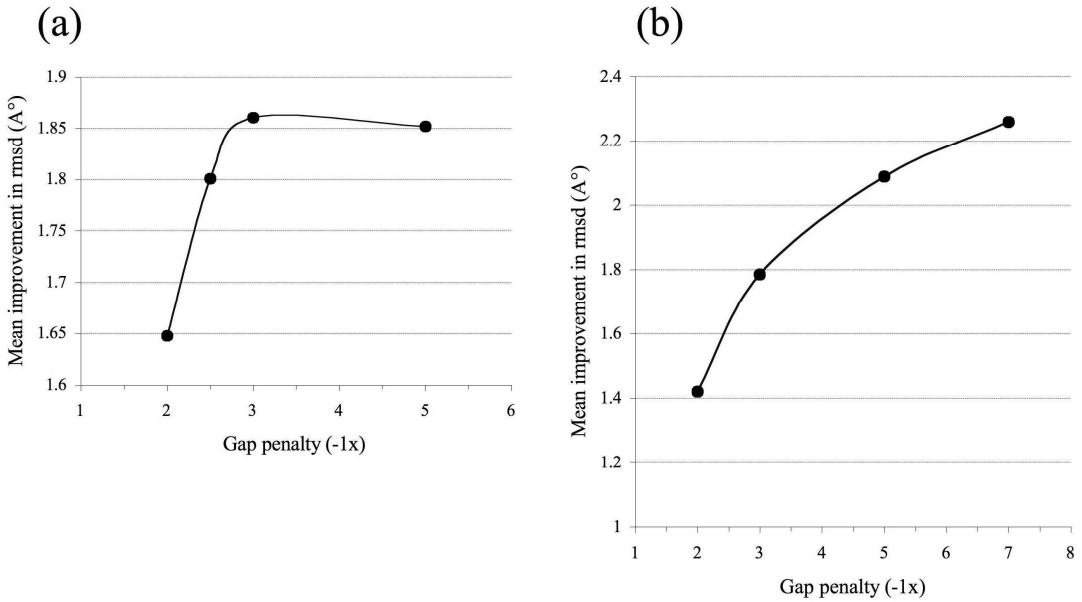
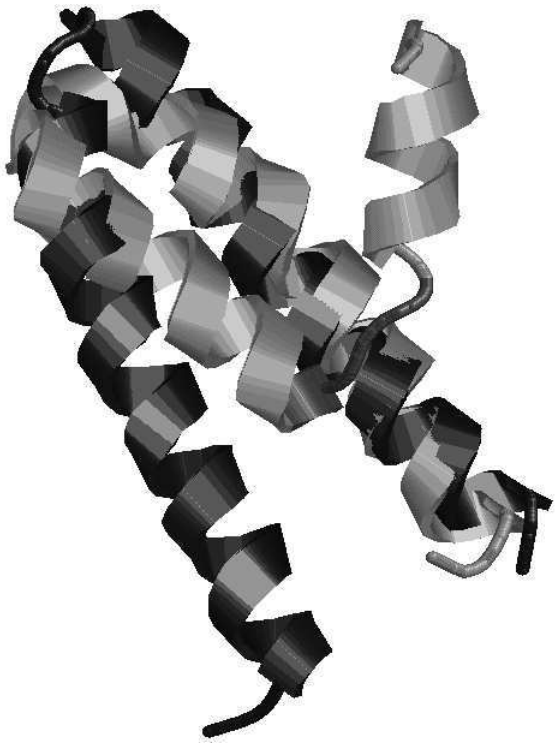


Figure 1. Effect of gap penalty on (a) global alignment and (b) local alignment.

Figure gives the mean improvement (decrease) in *rmsd* value (Y axis) at different negative gap penalties (X axis) with respect to *rmsd* values at gap penalty of –0.5. As shown, with increase in negative penalty there is an improvement in superimposed *rmsd* values compared to values obtained at penalty of –0.5. In case of local alignment (b), there is large improvement in alignment quality as negative gap penalty is increased. Even though –7.0 gives better mean improvement in *rmsd*, -5.0 was chosen as desired penalty as a balance between alignment quality and mining efficiency.



```
d1dd3a1#A   ZZKLMMMMMMMPCFKLMMMMMMMMMMMNOPKLMMMMMMMMMMMMMMMMMMMMNOZZ
d1b6q__#_   ZZKLMMMMMMMMMMMMMMMMMMMMMMMMPCFKLMMMMMMMMMMMMMMMMMMMM-ZZ
```

Figure 2. PB alignment based superimposition of SCOP domain d1dd3a1 (green) and d1b6q__ (blue).

PB alignment illustrates how long successions of PB *m* can contribute to alignment score. Presence of extra loop in protein 1DD3 is indicated in red color. It shows how small variation at local level can bring change in orientation of regular structures. Structural alphabet notation is explained in de Brevern et al[37,38].

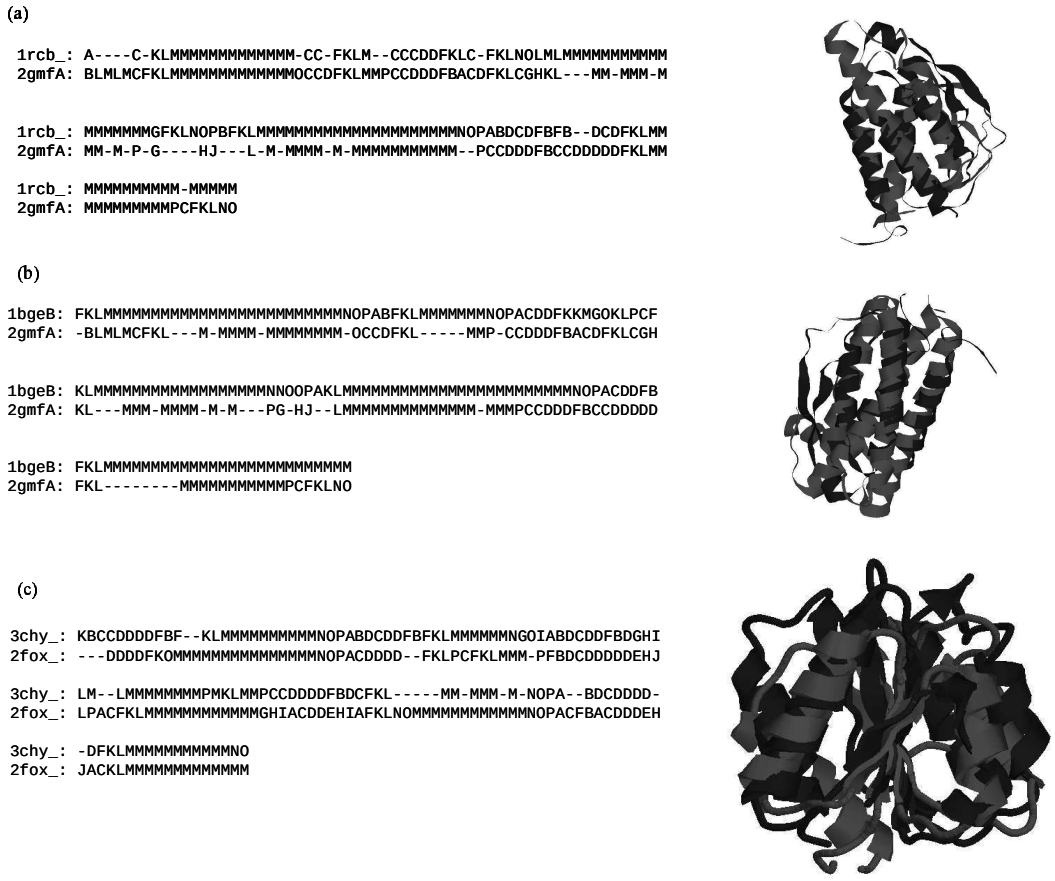


Figure 3. Global alignment of PB sequences.

(a) PB sequence alignment and superimposed structures for protein pair 1crb and 2gmfA. Target protein 2gmfA is found after using GA. (b) PB sequence alignment and superimposed structures for protein pair 1bgeB and 2gmfA. GA fails to find target protein 2gmfA. (c) Superimposed structures of 3chy and 2fox based on GA of PB sequences. Structural alphabet notation is explained in de Brevern et al[37,38].

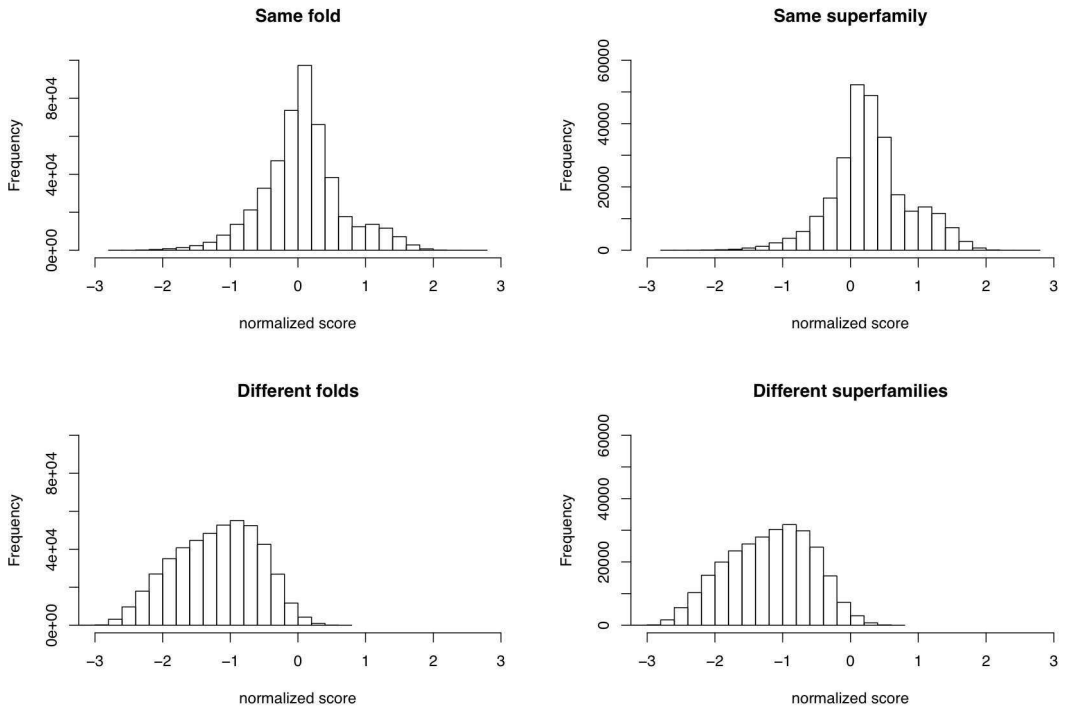


Figure 4. Distribution of normalized scores after PBs alignment between pairs of proteins from the same fold or superfamily (top) and from different folds or superfamilies (bottom).

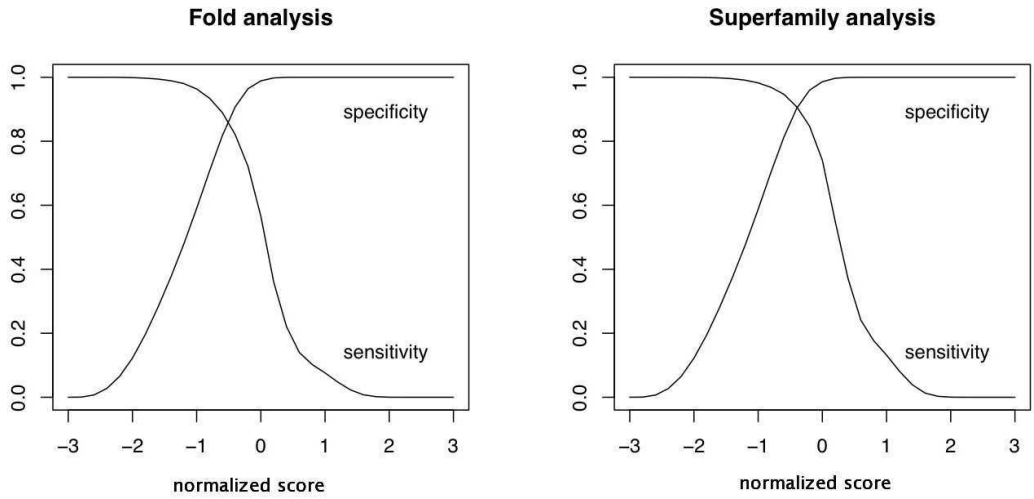


Figure 5. Analysis of variation of sensitivity and specificity according to different cut-off scores.

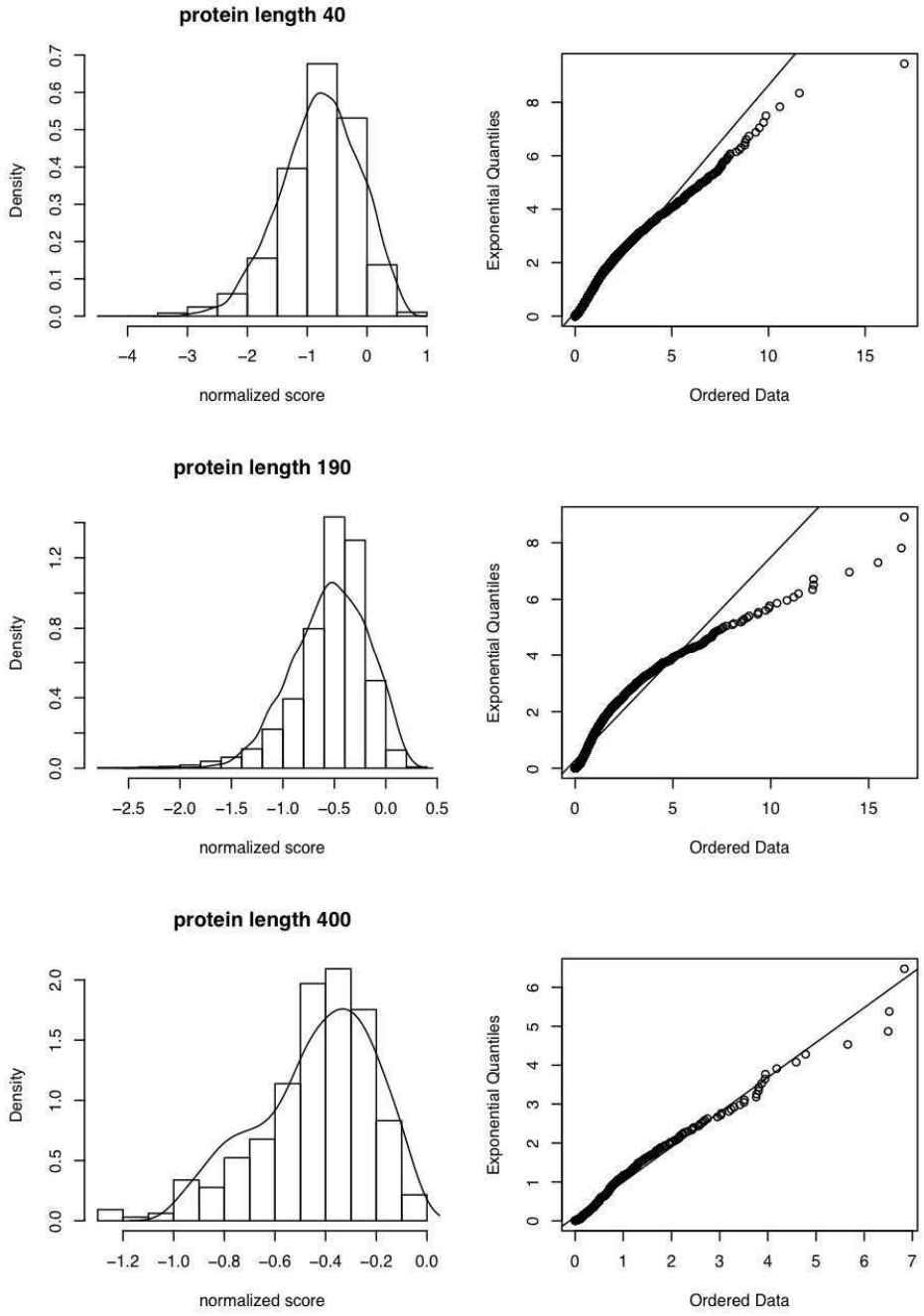


Figure 6. Distribution of scores from global alignments of real but unrelated sequences (RUS) datasets of 40aa, 190aa and 400aa long. The distribution of the scores was estimated with extreme value distribution curve indicated in solid line using *evir* package from R statistical software[51]. On the right are displayed the corresponding quantile plots

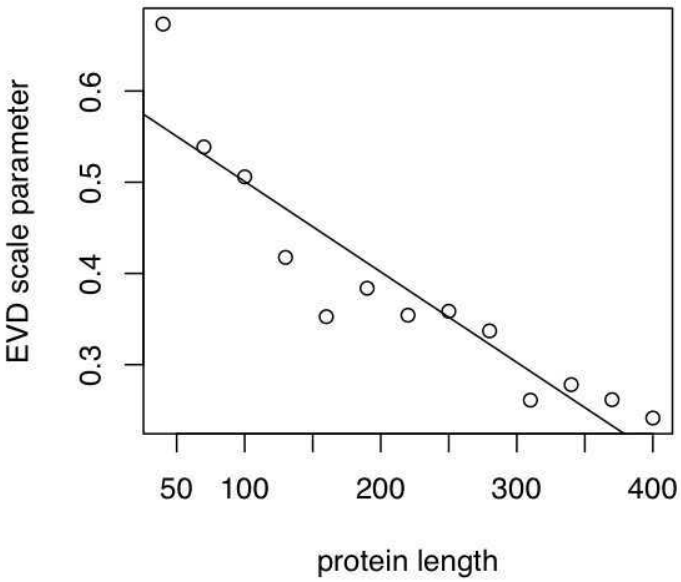


Figure 7. Variation of estimates of EVD scale parameter (σ) with length of protein sequences calculated from global alignments of real but unrelated sequences (RUS) datasets (see Table 11). Fitted regression line with R^2=0.85 ($p<0.0001$) as shown in solid line was calculated using *lm* function from R statistical software[51].

References

1. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. The Protein Data Bank. Nucleic Acids Res 2000;28(1):235-242.
2. Murzin AG, Brenner SE, Hubbard T, Chothia C. SCOP: a structural classification of proteins database for the investigation of sequences and structures. J Mol Biol 1995;247(4):536-540.
3. Novotny M, Madsen D, Kleywegt GJ. Evaluation of protein fold comparison servers. Proteins 2004;54(2):260-270.
4. Kolodny R, Koehl P, Levitt M. Comprehensive evaluation of protein structure alignment methods: scoring by geometric measures. J Mol Biol 2005;346(4):1173-1188.
5. Perutz MF. Structure of hemoglobin. Brookhaven Symp Biol 1960;13:165-183.
6. Orengo CA, Michie AD, Jones S, Jones DT, Swindells MB, Thornton JM. CATH--a hierarchic classification of protein domain structures. Structure 1997;5(8):1093-1108.
7. Shindyalov IN, Bourne PE. An alternative view of protein fold space. Proteins 2000;38(3):247-260.
8. Tyagi M, Sharma P, Swamy CS, Cadet F, Srinivasan N, Brevern AG, Offmann B. Protein Block Expert (PBE): A web-based protein structure analysis server using a structural alphabet. Nucl Acids Res 2006;34:W119-W123.
9. Holm L, Sander C. Protein structure comparison by alignment of distance matrices. J Mol Biol 1993;233(1):123-138.
10. Carpentier M, Brouillet S, Pothier J. YAKUSA: a fast structural database scanning method. Proteins 2005;61(1):137-151.
11. Krissinel E, Henrick K. Secondary-structure matching (SSM), a new tool for fast protein structure alignment in three dimensions. Acta Crystallogr D Biol Crystallogr 2004;60(Pt 12 Pt 1):2256-2268.
12. Blades MJ, Ison JC, Ranasinghe R, Findlay JBC. Automatic generation and evaluation of sparse protein signatures for families of protein structural domains. Protein Sci 2005;14(1):13-23.
13. Miguel RN. Sequence patterns derived from the automated prediction of functional residues in structurally-aligned homologous protein families. Bioinformatics 2004;20(15):2380-2389.
14. Domingues FS, Lackner P, Andreeva A, Sippl MJ. Structure-based evaluation of sequence comparison and fold recognition alignment accuracy. Journal of Molecular Biology 2000;297(4):1003-1013.
15. Sauder JM, Arthur JW, Dunbrack RL, Jr. Large-scale comparison of protein sequence alignment algorithms with structure alignments. Proteins 2000;40(1):6-22.
16. Friedberg I, Kaplan T, Margalit H. Evaluation of PSI-BLAST alignment accuracy in comparison to structural alignments. Protein Sci 2000;9(11):2278-2284.

17. Escalier V, Pothier J, Soldano H, Viari A. Pairwise and multiple identification of three-dimensional common substructures in proteins. J Comput Biol 1998;5(1):41-56.
18. Shindyalov IN, Bourne PE. Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. Protein Eng 1998;11(9):739-747.
19. Taylor WR, Orengo CA. Protein structure alignment. J Mol Biol 1989;208(1):1-22.
20. Levine M, Stuart D, Williams J. A method for the systematic comparison of the three-dimensional structures of proteins and some results. Acta Cryst 1984;40:600-610.
21. Usha R, Murthy MR. Protein structural homology: a metric approach. Int J Pept Protein Res 1986;28(4):364-369.
22. Singh AP, Brutlag DL. Hierarchical protein structure superposition using both secondary structure and atomic representations. Proc Int Conf Intell Syst Mol Biol 1997;5:284-293.
23. Lu G. TOP: a new method for protein structure comparisons and similarity searches. J Appl Crystallogr 2000;33:176-183.
24. Gibrat JF, Madej T, Bryant SH. Surprising similarities in structure comparison. Curr Opin Struct Biol 1996;6(3):377-385.
25. Harrison A, Pearl F, Mott R, Thornton J, Orengo C. Quantifying the similarities within fold space. J Mol Biol 2002;323(5):909-926.
26. Zuker M, Somorjai RL. The alignment of protein structures in three dimensions. Bull Math Biol 1989;51(1):55-78.
27. Subbiah S, Laurents DV, Levitt M. Structural similarity of DNA-binding domains of bacteriophage repressors and the globin core. Curr Biol 1993;3(3):141-148.
28. Vriend G. WHAT IF: a molecular modeling and drug design program. J Mol Graph 1990;8(1):52-56, 29.
29. Shatsky M, Nussinov R, Wolfson HJ. Flexible protein alignment and hinge detection. Proteins 2002;48(2):242-256.
30. Ye Y, Godzik A. Flexible structure alignment by chaining aligned fragment pairs allowing twists. Bioinformatics 2003;19 Suppl 2:II246-II255.
31. Offmann B, Tyagi M, de Brevern AG. Local protein structures. Curr Bioinf 2007;2(3):1-38.
32. Jones TA, Thirup S. Using known substructures in protein model building and crystallography. Embo J 1986;5(4):819-822.
33. Unger R, Harel D, Wherland S, Sussman JL. A 3D building blocks approach to analyzing and predicting structure of proteins. Proteins 1989;5(4):355-373.
34. Levitt M. Accurate modeling of protein conformation by automatic segment matching. J Mol Biol 1992;226(2):507-533.
35. Unger R, Sussman JL. The importance of short structural motifs in protein structure analysis. J Comput Aided Mol Des 1993;7(4):457-472.
36. Yang JM, Tung CH. Protein structure database search and evolutionary classification. Nucleic Acids Res 2006;34(13):3646-3659.
37. de Brevern AG. New assessment of a structural alphabet. In Silico Biol 2005;5(3):283-289.

38. de Brevern AG, Etchebest C, Hazout S. Bayesian probabilistic approach for predicting backbone structures in terms of protein blocks. Proteins 2000;41(3):271-287.
39. Tyagi M, Gowri VS, Srinivasan N, de Brevern AG, Offmann B. A substitution matrix for structural alphabet based on structural alignment of homologous proteins and its applications. Proteins 2006;65(1):32-39.
40. Dudev T, Lim C. Effect of carboxylate-binding mode on metal binding/selectivity and function in proteins. Acc Chem Res 2007;40(1):85-93.
41. Tyagi M, Sharma P, Swamy CS, Cadet F, Srinivasan N, de Brevern AG, Offmann B. Protein Block Expert (PBE): a web-based protein structure analysis server using a structural alphabet. Nucleic Acids Res 2006;34(Web Server issue):W119-123.
42. Smith TF, Waterman MS. Identification of common molecular subsequences. J Mol Biol 1981;147(1):195-197.
43. Needleman SB, Wunsch CD. A general method applicable to the search for similarities in the amino acid sequence of two proteins. J Mol Biol 1970;48(3):443-453.
44. Schuchhardt J, Schneider G, Reichelt J, Schomburg D, Wrede P. Local structural motifs of protein backbones are classified by self-organizing neural networks. Protein Eng 1996;9(10):833-842.
45. Tyagi M, Venkataraman SG, Srinivasan N, de Brevern AG, Offmann B. A substitution matrix for structural alphabet based on structural alignment of homologous proteins and its applications. Proteins in press.
46. Balaji S, Sujatha S, Kumar SS, Srinivasan N. PALI-a database of Phylogeny and ALIgnment of homologous protein structures. Nucleic Acids Res 2001;29(1):61-65.
47. Gowri VS, Pandit SB, Karthik PS, Srinivasan N, Balaji S. Integration of related sequences with protein three-dimensional structural families in an updated version of PALI database. Nucleic Acids Res 2003;31(1):486-488.
48. Russell RB, Barton GJ. Multiple protein sequence alignment from tertiary structure comparison: assignment of global and residue confidence levels. Proteins 1992;14(2):309-323.
49. McLachlan AD. Rapid Comparison of Protein Structres. Acta Cryst 1982;A38:871-873.
50. Lu G. TOP: a new method for protein structure comparisons and similarity search. J Appl Crystallogr 2000;33:176-183.
51. Ihaka R, Gentleman R. A language for data analysis and graphics. J Comput Graph Stat 1996;5:299-314.